KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy

L. Slaets, G. Claeskens and B. Silverman

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 0901

A Multiresolution Approach to Time Warping achieved by a Bayesian Prior-Posterior Transfer Fitting Strategy

Leen Slaets

Katholieke Universiteit Leuven, Belgium.

Gerda Claeskens

Katholieke Universiteit Leuven, Belgium.

Bernard Silverman

University of Oxford, United Kingdom.

Summary

The procedure known as warping aims at reducing phase variability in a sample of functional curve observations, by applying a smooth bijection to the argument of each of the functions. We propose a natural representation of warping functions in terms of a new type of elementary function named 'warping component functions' which are combined into the warping function by composition. A sequential Bayesian estimation strategy is introduced, which fits a series of models and transfers the posterior of the previous fit into the prior of the next fit. Model selection is based on a warping analogue to wavelet thresholding, combined with Bayesian inference.

Keywords: Bayesian Inference, Functional data analysis, Markov chain Monte Carlo sampling, Time warping, Warping components, Warping function.

1 Introduction

Functional data analysis refers to the statistical methodology designed for data sets involving functions, see Ramsay and Silverman (2006) for an explanation and overview. We restrict our attention to the setting of a sample of N curve observations, sampled from N unobserved continuous curves $(t, F_i(t))$ with domain $t \in [l, u]$, at discrete points t_{ij}, $i = 1, \ldots, N$. For clarity, we will refer to t as 'time' and the t_{ij} 'time points', although they do not have to represent time.

For many reasons, the rate at which data points are observed over time varies and does not necessarily reflect the underlying biological, physical, or other process governing the data. This is what we refer to as the phase variability in a sample of curves.

The other source of variability in a sample of functions is amplitude variability where there is variation in the function values themselves. It is important to recognize the identifiability problem between phase and amplitude variability (Ramsay and Silverman, 2006).

The current paper concentrates on the portion of the variability in the sample that can be attributed to the phase. Its presence might be of interest on its own or it can be a 'nuisance effect' that disturbs the analysis of the (possible) amplitude variability, see for instance Park (2008). Techniques that address phase variability are encountered by the terminology time warping, curve alignment, synchronization, or registration. Time warping is achieved by applying a transformation on the argument of the curves, the so-called warping function τ. By transforming the t coordinates by τ, the appearance of the new graph $\{(\tau(t), F_i(t))|t \in [l, u]\}$ will be different. In the one-dimensional case, in order to respect the natural ordering of the time points, the warping functions should be strictly monotone increasing to only delay or advance and to elongate or condense certain curve features. Continuity of a warping function avoids an infeasible split-up of the domain of the underlying smooth curve.

One possible approach is the landmark registration method set out, for example, by Kneip and Gasser (1992). It requires the manual identification of the location of a set of important features, or landmarks, often minima and maxima, for each of the sampled curves. These landmarks are then aligned to the average corresponding landmark by interpolating one-dimensional warping functions.

Continuous monotone registration is completely automatic and does not require the specification of landmarks nor the idea of interpolating a sequence of points. Silverman (1995) proposed this alignment by a time-shift with a Procrustes least squares estimation procedure and later Ramsay and Li (1996) incorporated flexible warping functions. The performance of this approach however, depends heavily on the unregistered cross-sectional average used in the Procrustes iteration. Other recent methods include Gervini and Gasser (2005), James (2007) and Telesca and Inoue (2008). They all achieve flexible warping by modelling the argument transformation as a regression or smoothing spline through a set of knots, with some type of monotonicity restriction. The spline basis expansion owes its success due to the vector space under addition and scalar multiplication formed by the target functions. The spline basis functions are, however, not warping functions themselves, which makes the components in the expansion not interpretable. Furthermore, the spline basis function approach implicitly considers the warping functions as members of the vector space of functions, subject to certain constraints.

This paper approaches warping transformations in a different way, as members of the group $\mathcal{W}_K$ of continuous transformations of the domain $[l, u]$ to itself. We introduce *elementary warping functions* or *warplets* denoted τ_i, warping surrogates for wavelets. They make a meaningful multiresolution analysis possible in the warping context. The natural way of combining warplets within the group

$\mathcal{W}_K$ is through composition instead of addition.

The main part of our study is devoted to a Bayesian estimation strategy, proposed in section 3. Telesca and Inoue (2008) also use a Bayesian warping approach, though with a hierarchical curve registration model and a penalized spline warping function. The Bayesian philosophy allows us to incorporate restrictions on parameters and to conduct exact inferences. Additionally our strategy fully exploits the opportunity to bring in prior information, by fitting a sequence of gradually extended models and each time transferring information from the posterior distribution to the prior in the next model. Bayesian inference on the warping parameters provides a natural model selection procedure, which resembles thresholding in a wavelet decomposition. A Markov chain Monte Carlo (MCMC) sampling scheme accounts for the computational aspect.

The performance and stability of the method is assessed by a simulation study in section 4, while section 5 contains an application to a proteomics data set (Listgarten et al., 2005). In section 6 we discuss possible extensions to more complex settings.

2 Warping Functions

By analogy to wavelet expansions we wish to decompose the warping function τ into components which are localized in location and scale. These building blocks, the *warping component functions* τ_i or warplets, are responsible for a local dilation and compression of a curve on a specific interval $[a-r, a+r]$ with center a and radius r. The intensity of the local deformation is governed by a third parameter λ. By composing several warplets into the warping function $\tau(t) = \tau_S \circ \ldots \circ \tau_2 \circ \tau_1(t)$, we can create a more complex deformation of the argument of the original curve. The resulting warping function consists of a number of effects localized in position and scale, which explains the terminology *multiresolution approach*.

2.1 Warping Components

Definition 2.1 formally defines the warping component function τ_i. It yields a continuous transformation which equals the identity transformation $I(t) = t$ outside the interval $[a-r, a+r]$.

This is achieved by placing a rescaled warplet kernel K function along the main diagonal (possibly reflected in the diagonal). Figure 1 illustrates this idea. A warplet τ_i induces a compression followed by a dilation, or the other way around depending on the sign of λ, and of a degree which is increasing with the absolute magnitude of λ. In order to maintain a bijective transformation, the absolute value

of parameter λ must not be too large. To be precise, assume that K is a function on $[-1,1]$ with $\sup|K'(t)| = 1$. We place the function $\lambda r\sqrt{2}K(rt\sqrt{2})$ on the diagonal centred at the point (a,a) as shown in the figure. Provided $-1 < \lambda < 1$, this construction will yield a monotone transformation of the interval $[a-r, a+r]$ to itself; the details of the transformation are set out in the definition.

The function K is in some ways similar to the kernel function in density estimation, though its key properties are different.

Definition 2.1 *The warping component function is defined as*

$$\tau_i\left((a,\lambda,r);x\right) = \begin{cases} a + rg\left(\lambda; \frac{(x-a)}{r}\right), & x \in [a-r, a+r] \\ x, & \text{otherwise,} \end{cases}$$

with $g(\lambda; y) = z + \lambda K(z)$ *in which* z *is the solution to*

$$z - \lambda K(z) = y, \tag{1}$$

with $\lambda \in (-1,1)$, $r > 0$ *and where the warplet kernel* K *is a symmetric continuous function on* $[-1,1]$, *and first derivative* $\sup_z |K^{'}(z)| = 1$.

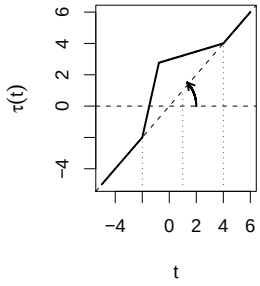


Figure 1: Warping component construction for $(a, r, \lambda) = (1, 3, 0.6)$ and the triangular warplet kernel $K^t(z)$ (Table 1).

It should be noted that it is a consequence of this definition that reversing the sign of the parameter λ yields the inverse transformation. Suppose $\tilde{y} = g(\lambda; y)$. Then $\tilde{y} = z + \lambda K(z)$ where $z - \lambda K(z) = y$. To find $g(-\lambda; \tilde{y})$, let $\tilde{\lambda} = -\lambda$. We then have z as the solution of $z - \tilde{\lambda}K(z) = \tilde{y}$, and so $g(\tilde{\lambda}, \tilde{y}) = z + \tilde{\lambda}K(z) = y$. Therefore $g(-\lambda; \tilde{y}) = y$. For the case of general a and r, substituting this result shows that $\tau((a,\lambda,r),.)$ and $\tau((a,-\lambda,r),.)$ are inverse transformations.

A variety of choices can be made for the warplet kernel. In particular, Table 1 provides the triangular function K^t, an Epanechnikov warplet kernel K^e and a quartic warplet kernel K^q, which are visualized in Figure 2. For all three exact solutions for (1) exist. The triangular and Epanechnikov warplet kernels lack smoothness. Only when using K^q the warping function will have a continuous first derivative over the entire domain. Therefore this is the preferred warplet kernel of the three when dealing with presmoothed curves $(t, F_i(t))$.

Table 1: Examples of warplet kernel functions K to use as building block for the warping components.

Warplet Kernel	Notation	Definition	c
Triangular	K^t	$K(z) = \begin{cases} 1-\lvert z\rvert, & z \in [-1,1] \\ 0, & \text{otherwise} \end{cases}$	1
Epanechnikov	K^e	$K(z) = \begin{cases} \frac{1}{2}(1-z^2), & z \in [-1,1] \\ 0, & \text{otherwise} \end{cases}$	$\frac{1}{2}$
Quartic	K^q	$K(z) = \begin{cases} \frac{3\sqrt{3}}{8}(1-z^2)^2, & z \in [-1,1] \\ 0, & \text{otherwise} \end{cases}$	$\frac{3\sqrt{3}}{8}$

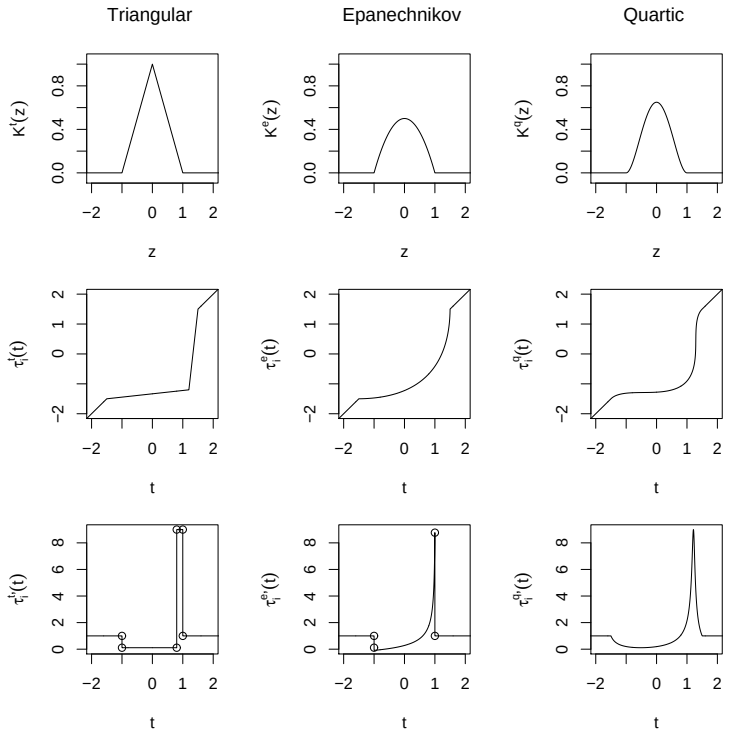


Figure 2: Kernel functions K to use as building blocks for the warping components (upper graphs), the corresponding warping components for $(a, \lambda, r) = (0, -0.8, 1.5)$ (middle graphs) and their first derivatives (lower graphs).

The first derivative of a warping component function at a certain point t is the rate in which an infinitesimal interval surrounding t ($\lim_{\varepsilon \downarrow 0}[t-\varepsilon, t+\varepsilon]$) will be dilated by the warping component. A compression corresponds to $\tau'(t) < 1$. Therefore it is called the instantaneous deformation rate. It characterizes each of the warplet kernels as shown in Figure 2 (lower panel). Irrespective of (a, r, λ),

K^t causes a steady deformation of the warping domain while for the other two the rate is not constant. The intensity λ of the warping component determines the extremes of τ_i', rather than simply the height of the rotated kernel function in τ_i, which should be taken relative to the interval $[a-r, a+r]$. Even though for a fixed λ the extremes are the same for each warplet kernel, see Figure 2 (lower panel), values of λ are not directly comparable across kernels. Because of the heights c of the warplet kernels (Figure 2, upper panel), the Epanechnikov and quartic functions require more extreme instantaneous deformation rates at certain time points compared to the triangular kernel for a similar warping action. This behaviour is needed to compensate lower deformation rates, enhancing the smoothness, around the center a for K^e and K^q or near the borders of $[a-r, a+r]$ for K^q. Section 4.1 contains a comparison of these kernels.

To increase flexibility we extend the symmetric warping components towards asymmetric ones, which are effective on intervals $[a-r_1, a+r_2]$.

Definition 2.2 *The asymmetric warping component function is defined as*

$$\tau_i\left((a,\lambda,r_1,r_2);x\right) = \begin{cases} a + r_1 \cdot g\left(\lambda\frac{r}{r_2}; (x-a)/r_1\right), & x \in [a-r_1, a-c\lambda r] \\ a + r_2 \cdot g\left(\lambda\frac{r}{r_2}; (x-a)/r_2\right), & x \in [a-c\lambda r, a+r_2] \\ x, & \text{otherwise,} \end{cases}$$

with $r_1, r_2 > 0$, $r = \min(r_1, r_2)$ *and* g, λ *as in Definition 2.1.*

The additional parameter r_2 in each component offers more flexibility and contributes to the parsimoniousness of the overall warping function. It remains the case that reversing the sign of λ yields the inverse transformation. Section 4.1 contains a comparative example to illustrate the advantage of using asymmetric warplets.

2.2 The Group of Warping Functions $\mathcal{W}_K$

Consider a finite number of warping component functions composed into a *warping function* τ, $\tau(t) = \tau_S \circ \ldots \circ \tau_2 \circ \tau_1(t)$. The collection of all such τ is named the *group of warping functions* $\mathcal{W}_K$ as described more formally in the following definition.

Definition 2.3 *We define the group of warping functions*

$$\begin{aligned} \mathcal{W}_{K,[l,u]} \;=\; & \{\tau : [l,u] \to [l,u] \;\mid\; \tau = \tau_n \circ \ldots \circ \tau_1, \textit{with component parameters} \\ & (a_i, \lambda_i, r_i)_{i=1,\ldots,n} \textit{ for which } [a_i - r_i, a_i + r_i] \subset (l,u), \textit{ for all } i, \textit{and where} \\ & \textit{each component has the same kernel } K\}. \end{aligned}$$

The next theorem states that $\mathcal{W}_{K,[l,u]}$ is a group. However, it is clear that it is not a commutative group. To facilitate notation, the subscript $[l, u]$ will often be abandoned in what follows. The proofs of all theorems in this section are gathered in Appendix A.1.

Theorem 2.1 *Warping functions form a group $\mathcal{W}_K$ under the composition operator.*

The group structure ensures that the inverse of a warping function is again a warping function. Moreover, the explicit formula of the inverse is easily obtained by changing the sign of the λ_i and reversing the order of the components,

$$(\tau_n(a_n, \lambda_n, r_n) \circ \ldots \circ \tau_1(a_1, \lambda_1, r_1))^{-1}(t) = \tau_1(a_1, -\lambda_1, r_1) \circ \ldots \circ \tau_n(a_n, -\lambda_n, r_n)(t). \tag{2}$$

This is a particular advantage of the warplets compared to other warping methods, as we will show in section 3.1. In the absence of a vector space and because a decomposition into warping components is never unique, the terminology basis function is avoided and replaced with *elementary warping functions* or *warping components* and *dictionary* of warping components. Theorem A.1 in the appendix states that a warping function τ can always warp some arbitrary x and y values in (l, u) to each other, while not disturbing any other value outside $[x, y]$ or $[y, x]$. The most important conclusion that we obtain in Theorem A.2 (6) in the appendix, is that we can approximate every strictly monotone increasing (s.m.i) surjective continuous transformation on $[l, u]$ arbitrarily close by an element $\tau \in \mathcal{W}_K$. Further research is required to learn more about the quality of the approximation, as implied by this denseness.

3 A Bayesian, Prior-Posterior Transfer Estimation and Model Selection Strategy

We adopt the Bayesian setting in which information coming from the data and represented by the likelihood is combined with prior believes concerning the parameters in the model. The possibility of including prior information is in particular exploited for the construction of a special model fitting and selection strategy.

3.1 Model Formulation and Data Likelihood

The data are realizations $y_i(t)$ of random variables $Y_i(t)$, that are noisy versions of N unobserved continuous curves $(t, F_i(t))$ with domain $[l, u]$ for $i = 1, \ldots, N$, and are only observed at discrete time points t_{ij}, $j = 1, \ldots, n_i$,

$$Y_i(t_{ij}) = F_i(t_{ij}) + \varepsilon_{ij}, \quad \text{with} \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, N \text{ and } j = 1, \ldots, n_i, \tag{3}$$

where we assume independent, identically distributed (i.i.d.) normal errors.

First, we introduce the method for a sample of two curves, that are warped versions of each other,

$$(m(t), F_2(t)) = (t, F_1(t)) \text{ or } (t, F_2(t)) = (m^{-1}(t), F_1(t)) \text{ or } (t, F_2(t)) = (t, F_1(m(t))),$$

with $m : [l, u] \to [l, u]$ s.m.i, surjective and continuous. We estimate m by a warping function τ in $\mathcal{W}_K$ and use a fixed number of asymmetric warping components S_i for each curve. See section 3.3 for incorporating the selection of the components. Because of (2) we can treat the following models equally,

$$Y_1(t_{1j}) = Y_2(\tau^{-1}(t_{1j})) + \xi_{1j}, \quad Y_2(t_{2j}) = Y_1(\tau(t_{2j})) + \xi_{2j}, \quad \text{with } \xi_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 2\sigma^2), \tag{4}$$

for $i = 1, 2,\ j = 1, \ldots, n_i$, instead of choosing one curve as a reference. It is clear that this is an important asset of the warping components. The models in (4) require the estimation of only one warping function. More generally, we can warp both curves, with constraints to avoid over-identifiability. Indeed, consider an arbitrary warping function τ_1 in $\mathcal{W}_{K,[l,u]}$. Since $F_1(\tau_1(t)) = F_2(\tau^{-1}(\tau_1(t))$, a more general model formulation would be

$$Y_1(\tau_1(t_{1j})) = Y_2(\tau_2(t_{1j})) + \xi_{1j}, \quad Y_2(\tau_2(t_{2j})) = Y_1(\tau_1(t_{2j})) + \xi_{2j}, \quad \text{with } \xi_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 2\sigma^2),$$

for $i = 1, 2,\ j = 1, \ldots, n_i$, where in (4) the constraint $\tau_1 = I$ is applied. Other constraints can be chosen to reparameterize the model, like for instance $\tau_1 \circ \tau_2 = I$, which means that we warp both curves to some unobserved function located in the 'middle', a 'horizontal average'.

The following log-likelihood functions are used to estimate the warping function parameters in τ and the variance σ^2, corresponding to models (4),

$$\begin{aligned}
\log L_1(\boldsymbol{\alpha}, \sigma^2) &= -n_1 \log(2\sigma\sqrt{\pi}) - \sum_{j=1}^{n_1} \left\{ y_1(t_{1j}) - y_2\left(\tau^{-1}(t_{1j})\right) \right\}^2 / (4\sigma^2), \\
\log L_2(\boldsymbol{\alpha}, \sigma^2) &= -n_2 \log(2\sigma\sqrt{\pi}) - \sum_{j=1}^{n_2} \left\{ y_2(t_{2j}) - y_1\left(\tau(t_{2j})\right) \right\}^2 / (4\sigma^2),
\end{aligned}$$

with $\boldsymbol{\alpha}$ the collection of the unknown warping parameters $\{a_s, \lambda_s, r_{1,s}, r_{2,s}\}_{s=1}^{S}$ for S asymmetrical components. To account for both models equally, we use a weighted log-likelihood

$$\begin{aligned}
\log L(\boldsymbol{\alpha}, \sigma^2) &= -\frac{2n_1 n_2}{n_1 + n_2} \log(2\sigma\sqrt{\pi}) \\
&- \left[\frac{n_2}{n_1 + n_2} \sum_{j=1}^{n_1} \left\{ y_1(t_{1j}) - y_2\left(\tau^{-1}(t_{1j})\right) \right\}^2 + \frac{n_1}{n_1 + n_2} \sum_{j=1}^{n_2} \left\{ y_2(t_{2j}) - y_1\left(\tau(t_{2j})\right) \right\}^2 \right] / (4\sigma^2).
\end{aligned}$$

For a sample of $N \geq 2$ curves, this is generalized towards

$$L(\boldsymbol{\alpha}, \sigma^2) = (2\sigma\sqrt{\pi})^{-\frac{N(N-1)\prod_{i=1}^{N} n_i}{(N-1)\sum_{i=1}^{N}(\prod_{k\neq i} n_k)}} \times \exp\left[-\sum_{i=1}^{N}\left[\frac{\prod_{k\neq i} n_k}{(N-1)\sum_{i=1}^{N}(\prod_{k\neq i} n_k)}\sum_{k\neq i}\left(\sum_{j=1}^{n_i}\left\{y_i(\tau_i(t_{ij})) - y_k\left(\tau_k(t_{ij})\right)\right\}^2\right)\right]/(4\sigma^2)\right], \tag{5}$$

in which $\boldsymbol{\alpha} = \{a_{i,s}, \lambda_{i,s}, r_{1,i,s}, r_{2,i,s}\}_{s=1\ldots S_i}^{i=1\ldots N}$, and the constraint $\tau_i = I$ for a certain i could be applied. We remark that the $y_i(\tau_i^{-1}(t_{ij}))$ are most likely not observed. Since the τ_i belong to $\mathcal{W}_{K,[l,u]}$ and assuming that the time points t_{ij} are sufficiently well spread over the time domain, interpolation or prediction by a smoothing method can overcome this issue (see section 3.3). From now on a different parameterization will be used, in which the warping lower and upper bounds $w_l = a - r_1$ and $w_u = a + r_2$ replace the two radii.

3.2 MCMC posterior sampling

We take the following non-informative priors for the model parameters

$$\begin{aligned} a_s^i &\sim U(l,u),\ w_{l,i,s} \sim U(l,u),\ w_{u,i,s} \sim U(l,u),\ \lambda_s^i \sim U(-1,1) \\ \sigma^2 &\sim IG(\epsilon,\epsilon), \quad \text{with } \epsilon \text{ very small}, \quad s = 1,\ldots,S_i \text{ and } i = 1,\ldots,N. \end{aligned} \tag{6}$$

$U(x_1, x_2)$ denotes the uniform distribution on the interval (x_1, x_2) and $IG(x_1, x_2)$ the inverse gamma distribution with scale x_1 and shape parameter x_2. The prior for σ^2 is the proper alternative for Jeffrey's improper prior and for ϵ sufficiently small, the prior impact is negligible.

The aim is to obtain samples from the posterior distribution corresponding to the priors in (6), with data prior $f\left(\boldsymbol{\alpha}, \sigma^2\right)$, and to the likelihood $f\left(\{y_i(t_{ij})\}_{j=1\ldots n_i}^{i=1\ldots N}|\boldsymbol{\alpha}, \sigma^2\right) = L(\boldsymbol{\alpha}, \sigma^2)$ in (5),

$$f_{post}\left(\boldsymbol{\alpha}, \sigma^2\right) = f\left(\boldsymbol{\alpha}, \sigma^2 \middle| \{y_i(t_{ij})\}_{j=1\ldots n_i}^{i=1\ldots N}\right) = \frac{f\left(\{y_i(t_{ij})\}_{j=1\ldots n_i}^{i=1\ldots N}|\boldsymbol{\alpha}, \sigma^2\right) f\left(\boldsymbol{\alpha}, \sigma^2\right)}{\int f\left(\{y_i(t_{ij})\}_{j=1\ldots n_i}^{i=1\ldots N}|\boldsymbol{\alpha}, \sigma^2\right) f\left(\boldsymbol{\alpha}, \sigma^2\right) d(\boldsymbol{\alpha}, \sigma^2)}. \tag{7}$$

The Metropolis-Hastings algorithm (see for instance Chib and Greenberg, 1995), which is an iterative Markov chain Monte Carlo (MCMC) procedure, requires the specification of a so-called proposal density, to generate a proposal sample. The latter will then either be retained or rejected. In order for the algorithm to converge sufficiently fast, the acceptance rate should roughly be in between 20%–40%. It is not immediately clear how the variability of the proposal density relates to the acceptance rate, which is, moreover, not constant during the procedure due to the burn-in period. It is therefore advisable to allow for regular updates of the proposal density variance. An MCMC algorithm is well known to generate dependent samples, leading to high serial correlation in the chains. To augment

their information content while not increasing computational memory, iterated parameter values are only stored on a regular basis, which is called thinning.

A starting value $\{a_{i,s}^{(1)}, \lambda_{i,s}^{(1)}, w_{l,i,s}^{(1)}, w_{u,i,s}^{(1)}, \sigma^{2(1)}\}_{s=1\ldots S_i}^{i=1\ldots N}$ is determined in accordance with the prior distribution and the procedure generates new proposals

$$\{\boldsymbol{\alpha}^{(p+1)}, \sigma^{2(p+1)}\} = \{a_{i,s}^{(p+1)}, \lambda_{i,s}^{(p+1)}, w_{l,i,s}^{(p+1)}, w_{u,i,s}^{(p+1)}, \sigma^{2(p+1)}\}_{s=1\ldots S_i}^{i=1\ldots N}$$

based on the previous accepted values $\{\boldsymbol{\alpha}^{(p)}, \sigma^{2(p)}\} = \{a_{i,s}^{(p)}, \lambda_{i,s}^{(p)}, w_{l,i,s}^{(p)}, w_{u,i,s}^{(p)}, \sigma^{2(p)}\}_{s=1\ldots S_i}^{i=1\ldots N}$ by the proposal density. The latter is denoted by q and equals the product of the densities of the following distributions

$$\begin{aligned}
\lambda_{i,s}^{(p+1)} \quad &\text{drawn from} \quad \bar{\mathcal{N}}\left(\lambda_{i,s}^{(p)}\ ,\ \sigma_p^2, -1, 1\right) \\
w_{u,i,s}^{(p+1)} \quad &\text{drawn from} \quad \bar{\mathcal{N}}\left(w_{u,i,s}^{(p)}\ ,\ \sigma_p^2 \cdot \frac{u-l}{2}, l, u\right) \\
w_{l,i,s}^{(p+1)} \quad &\text{drawn from} \quad \bar{\mathcal{N}}\left(w_{l,i,s}^{(p)}\ ,\ \sigma_p^2 \cdot \frac{u-l}{2}, l, w_{u,i,s}^{(p+1)}\right) \\
a_{i,s}^{(p+1)} \quad &\text{drawn from} \quad \bar{\mathcal{N}}\left(a_{i,s}^{(p)}\ ,\ \sigma_p^2 \cdot \frac{u-l}{2}, w_{l,i,s}^{(p+1)}, w_{u,i,s}^{(p+1)}\right) \\
\sigma^{2(p+1)} \quad &\text{drawn from} \quad \bar{\mathcal{N}}\left(\sigma^{2(p)}\ ,\ \sigma_p^2 \cdot v\right),
\end{aligned} \tag{8}$$

in which σ_p^2 stands for proposal variance and $\bar{\mathcal{N}}(x_1, x_2, x_3, x_4)$ denotes the truncated normal distribution on the interval (x_3, x_4) with mean x_1 and variance x_2. The variances of the proposal densities equal σ_p^2 multiplied by a certain constant to account for the corresponding parameter range. An initial guess $\sigma_p^2 \cdot v$ is used as the proposal variance for σ^2. We advise a relatively large initial choice for the proposal variance (e.g. 0.8) to benefit the exploration phase. Throughout the algorithm, this will be updated to better suit the posterior distribution. The truncated normal distribution makes sure that the generated warping parameter proposals indeed give rise to a warping function in $\mathcal{W}_{K,[l,u]}$. The new proposal is evaluated by the Metropolis-Hastings algorithm and approved with probability

$$P(\text{acceptance}) = \min\left\{ \frac{f_{post}\left(\boldsymbol{\alpha}^{(p+1)}, \sigma^{2(p+1)}\right)}{f_{post}\left(\boldsymbol{\alpha}^{(p)}, \sigma^{2(p)}\right)} \cdot \frac{q\left(\boldsymbol{\alpha}^{(p)}, \sigma^{2(p)} | \boldsymbol{\alpha}^{(p+1)}, \sigma^{2(p+1)}\right)}{q\left(\boldsymbol{\alpha}^{(p+1)}, \sigma^{2(p+1)} | \boldsymbol{\alpha}^{(p)}, \sigma^{2(p)}\right)}, 1 \right\}. \tag{9}$$

If the proposed value is rejected, the previous one will be carried over.

When using this acceptance–rejection routine we obtain a sample of the posterior distribution. The latter is often visualized by means of histograms for each of the model parameters. An important advantage of this Bayesian setting is the availability of a complete sample of the joint posterior distribution of the model parameters, which allows for exact inferences.

3.3 Prior-Posterior Transfer and Model Selection

The previous sections offer all the necessary building blocks of the Metropolis-Hastings algorithm. There are, however, some important remaining issues. First of all, since a decomposition into components is not necessarily unique and unneeded components could be present, various combinations lead to quasi identical fits. Consequently, multimodal posterior distributions might occur when estimating a model with more than one component. This is undesirable since the components hold no clear interpretation and combining the parameters from many marginal multimodal posterior distributions forms quite a challenge. Further, how many warping components should be considered for each warping function?

These two remarks motivate the use of the following estimation strategy. We propose to exploit the multiresolution structure by condensing the necessary warping actions in as little components as possible and delete all the sparse ones. By sparse components we mean components with a relatively narrow warping domain $[w_{l,s}, w_{u,s}]$ or a small intensity λ_s. The elimination based on λ resembles that of wavelet thresholding, while a small warping domain is related to detailed high frequency information. To achieve these goals, a sequence of models is fitted, each time adding a warping component for which the posterior of the previous fit is transferred to the prior of the next, as proposed in the following strategy.

1. First perform the Bayesian estimation method as in section 3.2 with the warping functions τ_i consisting of one component. This results in a sample from the posterior distribution of the parameters $(a_{i,1}, \lambda_{i,1}, w_{l,i,1}, w_{u,i,1})$ of the single warping components which account for the most important warping action. To predict the unobserved $y_i(\tau_i^{-1}(t_{ij}))$ values, two methods are implemented in our `R` program: simple linear interpolation and prediction of the values by means of a penalized spline fit to the curve based on the observed data points (`spm` in the `R`-package `SemiPar`). These predicted values display less variability than the original data points $y_i(t_{ij})$. As a result the error variance in (4) is smaller and lies in between σ^2 and $2\sigma^2$.

2. In the second step we add one component to the warping function, to eliminate as much as possible of the remaining phase variability after warping by the first component. This is achieved by updating the prior (6) on the parameters of the first components $(a_{i,1}, \lambda_{i,1}, w_{l,i,1}, w_{u,i,1})$ by the posterior distribution as obtained in step 1. In practice four marginal histograms for each i, with a fixed number of bins, serve as a crude estimate of this $(4i)$-dimensional posterior distribution. The proposal density (8) is updated for the first warping component parameters by replacing

$(u - l)$ by the difference between the upper and lower bounds of their new priors.

3. We continue by step by step adjusting the priors and proposal density of the previous components while adding a new one. By updating the priors of the component parameters it follows that the necessary warping actions are condensed into the first components in each fitted model, rather than spread out over all the available components. This largely avoids multimodel posterior distributions and makes sure that later components contribute less to the overall warp.

4. A natural model selection approach arises. At a certain point the newest component s does not perform better or performs even worse than the component with $\lambda_{i,s} = 0$ or $r_{1,i,s} = 0 = r_{2,i,s}$ (empty component). This means that zero values for the latter component parameters are likely. Based on the sample of the posterior distribution of the newest component, marginal $(1 - \alpha)$ highest posterior density (hpd) intervals for λ_s, $w_{l,s}$ and $w_{u,s}$ are computed by the function `emp.hpd` of the `R`-package `TeachingDemos`. We then delete the latest component s and stop the fitting procedure when zero is contained within the hpd-interval for λ_s or when the lower bound of the hpd-interval for $w_{u,s}$ is lower than the upper limit of the hpd-interval for $w_{l,s}$.

In our `R` program we implemented the two stopping criteria for the addition of components, as described in step 4. We used the starting value $\lambda_{s+1} = 0.0001$ for the relative intensity of each new component (the value of zero causes numerical problems). The new component is only included when it performs sufficiently better than no warp at all.

4 Simulation Study

4.1 Illustrative example with comparison of warplet kernels

Figure 3 (a) displays two curve observations, in which the curve with the bold dots $(t, F_2(t))$ is a warped version $(m(t), F_1(t))$ of the other curve $(t, F_1(t))$. The observations contain 200 time points t_{ij} each, with corresponding function values $y_i(t_{ij})$ and constitute a smooth curve plus some random $\mathcal{N}(0, 0.16)$ noise. The true argument transformation is shown in the same figure in the right panel.

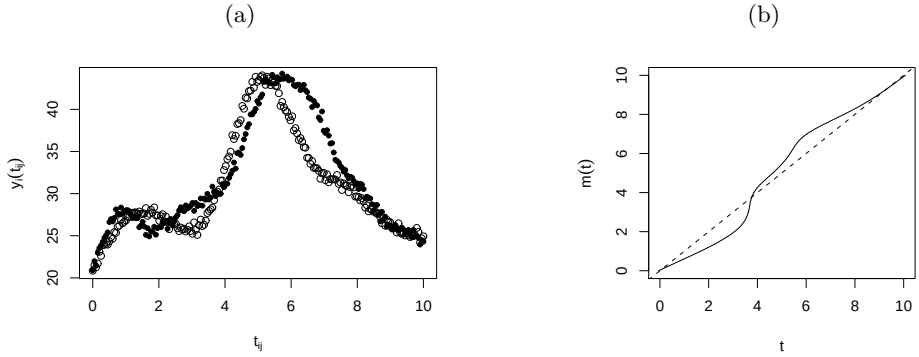


Figure 3: A misaligned curve sample (a) and the true unobserved argument transformation (b).

We use our estimation routine (with penalized spline predicted function values for evaluation of the likelihood) to estimate a warping function τ as approximation of m^{-1}, which synchronizes the curves. We first use the quartic warplet kernel, later we compare the results with those for the other warplet kernels and linear interpolation predictions in the likelihood.

In each fitted model the MCMC sampling procedure generated 2000 samples from the posterior distribution of the model parameters. They resulted from an original sample in which thinning was applied to store only every $(30 + 2S)$th generated set of values, with S the number of warping components in the model. The first 2000 retained samples were removed to exclude the burn-in period. In this example, we took $\left(\epsilon, \sigma_p^2, v, \alpha\right) = (0.000001, 0.8, 0.1, 0.95)$. Figure 4 (first row) shows histograms of the sampled warping parameters for the first model with only one component. The set of values corresponding to the highest likelihood value (point estimates) are marked by a vertical line and the hpd bounds for λ_1, $w_{l,1}$ and $w_{u,1}$ by bold vertical lines. The parameter histograms show that λ_1 is not likely to be zero nor are the lower and upper limit histograms overlapping. This justifies the inclusion of this warping component and the allowance for an additional component in the next fit. These histograms serve as priors on the parameters of the first warping component in the new model with two components. The point estimates are always used to plot the current warped curves (see Figure 5) and estimated warping function τ^{-1} (Figure 6). In the extended model with two components the second component is still able to eliminate some remaining phase variability (Figure 5). The non-redundancy of this component is clearly confirmed by the histograms in Figure 4.

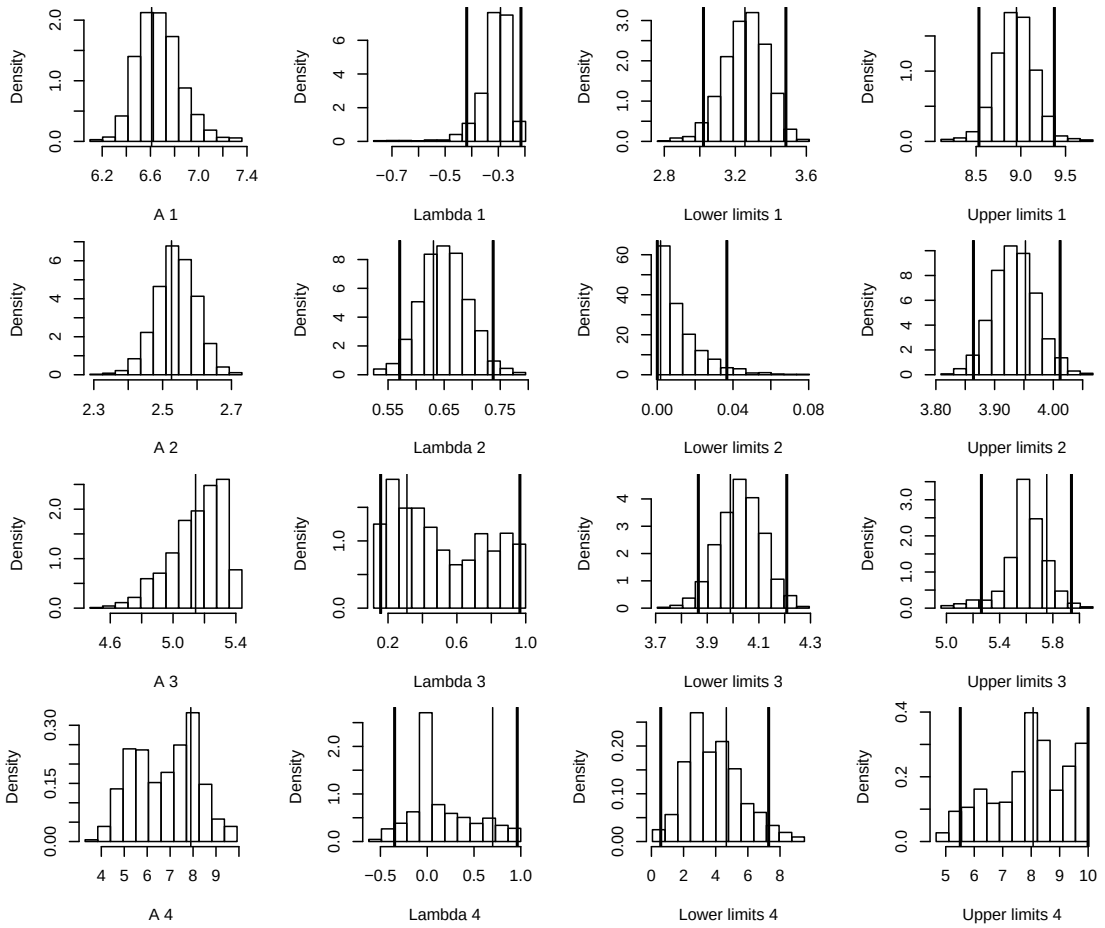


Figure 4: In row J, histograms of sampled warping parameters a_{J+2}, λ_{J+2}, $w_{l,J+2}$ and $w_{u,J+2}$ in the model with J+2 warping components.

We continue with temporarily allowing a third component, where the posteriors for the first two components are transferred to the prior in this new model. The third component still sufficiently aids in the alignment of the highest peak of the curve observations (Figure 4 and 5). For the fourth component the situation is remarkably different. No clear visual improvement was achieved compared to the previous model (Figure 5). The histogram for λ_4 reveals that the addition of the fourth component does not perform noticeably better than the addition of a component with a zero relative intensity. We conclude that the warping function with three components as plotted in Figure 6 (lower left panel) is suitable to describe the domain transformation.

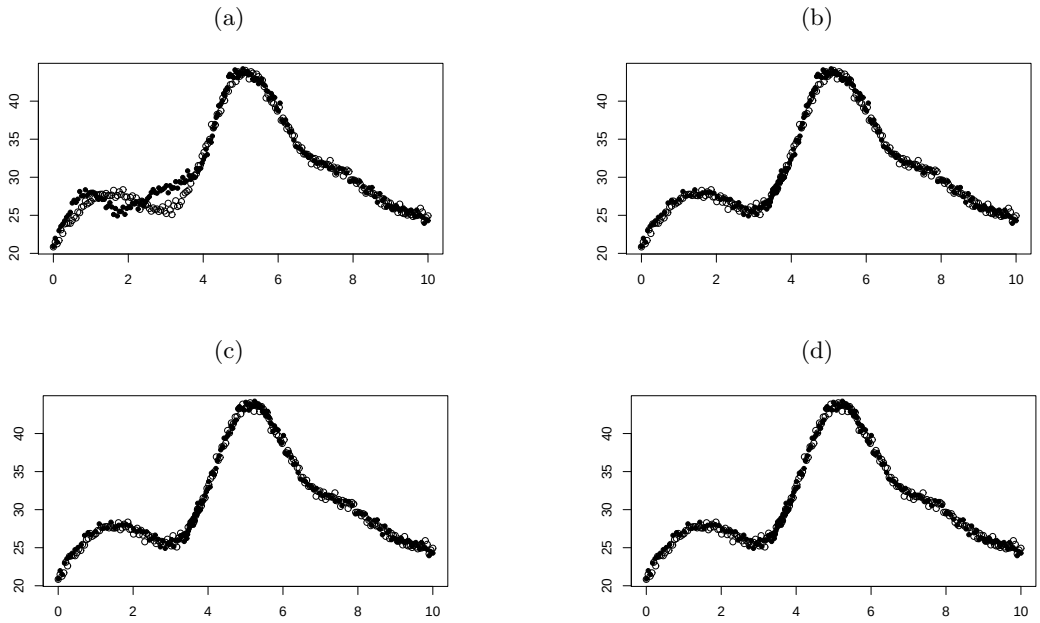


Figure 5: Plot of the warped curves for the model with (a) one component, (b) two, (c) three and (d) four components.

We compare these results with registration using the Epanechnikov and triangular warplet kernels. Figure 7 contains the warped curves and estimated warping functions, while Table 2 provides some summary information regarding the fit. The root average square error (RASE) describes the amplitude variability between the warped curve observations $\{\tau(t_{2j}), y_2(t_{2j})\}_{j=1,\ldots,200}$ and the true underlying smooth curve F_1,

$$\text{RASE} = \sqrt{\frac{1}{n_2}\sum_{j=1}^{n_2}\{y_2(t_{2j}) - F_1(\tau(t_{2j}))\}^2}.$$

For a decent warp it should be close to the standard deviation of the error term (0.4 in this case). Based on the RASE, the triangular warplet kernel results in the most effective warp, followed by the quartic warplet kernel. The 'best' warplet kernel to model a particular warp merely depends on the underlying unobserved transformation m and the shape of the curves. The quartic function is the only one respecting the smooth nature of the curves and is the most 'natural' looking. If the data will still undergo a smoothing stage after the alignment and are recorded with a reasonable error, we can adopt the triangular warping components, with the advantage of decreased computation time. Moreover, the triangular warplets strike the golden mean between the concave and convex transitions towards the exterior of the warping domain of respectively the Epanechnikov and quartic warplets. This might result in the best fit for transformations with both concave and convex features and not the worst fit for a transformation m with just one of the two features. A post-smoothing procedure and/or highly variable error term would counteract the sharp edges of the warping function.

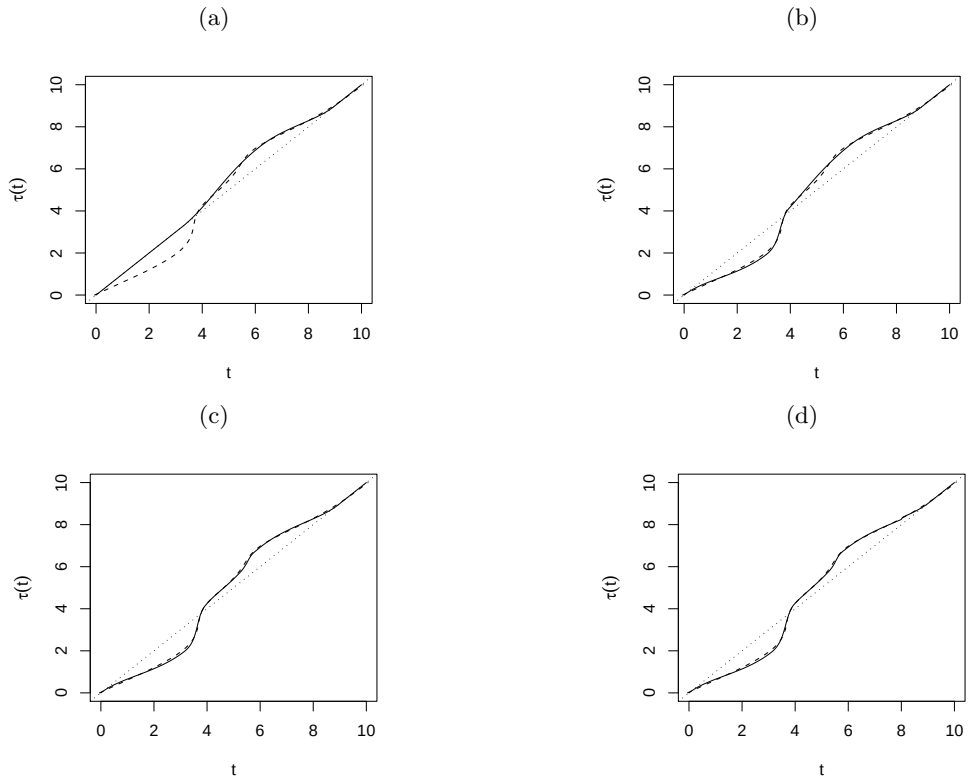


Figure 6: Plot of the estimated warping function (solid line) together with the true transformation (dashed line) for the model with (a) one component, (b) two, (c) three and (d) four components.

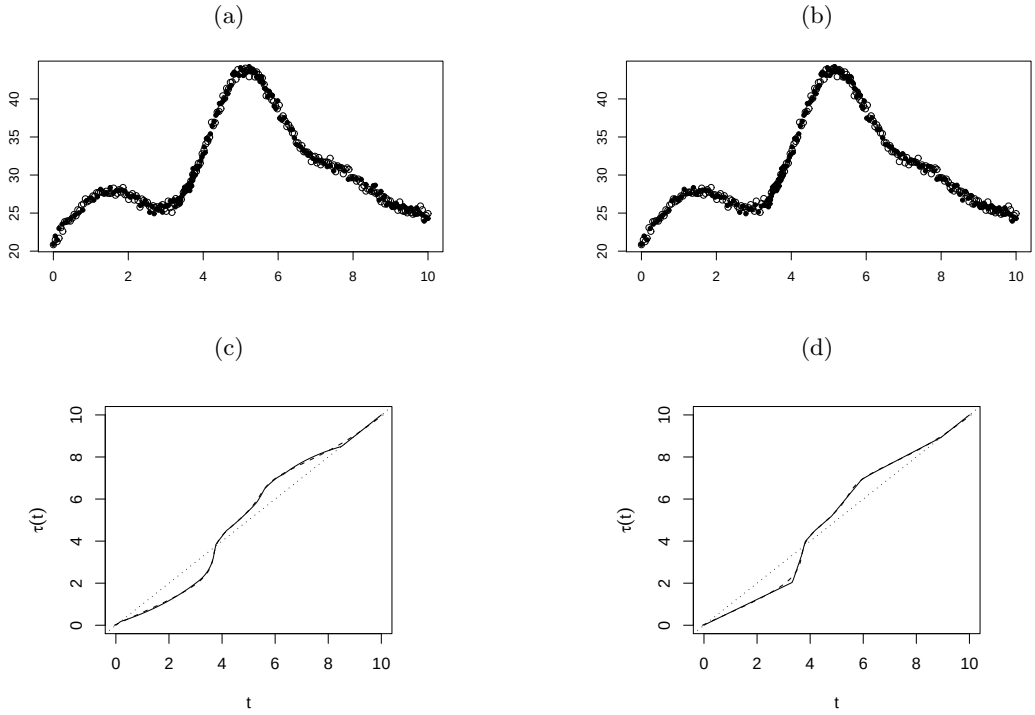


Figure 7: Plot of the warped curves and estimated warping functions for the Epanechnikov kernel (a) and (c) and the triangular kernel (b) and (d).

Three additional models were fitted to the curves, in which we repeated the above method with

the linear interpolation function value predictions in the likelihood. The fits are slightly better with the penalized spline predictions and more stable concerning the elapsed time. Even though the run for a single fit is faster with linear interpolation, it selected more components for K^e and K^t. In the next section we will evaluate whether these remarks hold in general. A simulation study is set up to assess the stability and performance of the method for both predictions in the likelihood.

Table 2: Performance comparison for the quartic, Epanechnikov and triangular warplet kernels and for the linear interpolation and penalized spline function value predictions.

Kernel	Linear interpolation predictions			Penalized spline predictions		
	RASE	♯ components	elapsed time	RASE	♯ components	elapsed time
K^q	0.4297	3	1:30	0.4137	3	2:32
K^e	0.4252	6	4:05	0.4242	4	3:28
K^t	0.4107	4	2:04	0.3972	3	2:23

Finally we illustrate for this example the advantage of using asymmetric components over symmetric ones. Exactly the same model with the penalized spline predictions and the quartic warplet kernel is fitted, but now with only symmetric components. As a result, 6 components were retained for a total of 18 warping parameters. The computational time more than doubled (5:35) as compared to the asymmetric components, for a final RASE of 0.3974.

4.2 Likelihood evaluation based on linear interpolation versus penalized spline function value predictions

In this simulation study, the experimental units are generated according to model (3), with the underlying smooth curve F_1 as in (10),

$$\begin{aligned} F_1(t) &= \frac{70}{1.1\sqrt{2\pi}}\exp\left(-\frac{(x-5)^2}{2}\right)+\frac{120}{1.6\sqrt{2\pi}}\exp\left(-\frac{(x-2.5)^2}{8}\right) \\ &\quad +\frac{100}{1.2\sqrt{2\pi}}\exp\left(-\frac{(x-7.5)^2}{5.12}\right)+\frac{75}{1.1\sqrt{2\pi}}\exp\left(-\frac{(x-10.6)^2}{2.88}\right). \end{aligned} \tag{10}$$

The argument transformation m is now a warping function itself, consisting of two triangular components with parameters $(a_1, \lambda_1, r_{11}, r_{21}) = (2.5, 0.17, 1.8, 3)$ and $(a_2, \lambda_2, r_{12}, r_{22}) = (6, -0.2, 2.13, 3.6)$. Four different error variances, 0.2^2, 0.5^2, 0.8^2 and 1 are considered. With these four settings we wish to investigate whether for relatively large variances, the penalized spline function value predictions result in more stable and accurate time warpings, while for smaller variances the difference between

the two methods will be less apparent. We used the kernel K^t and the same algorithm settings as in the previous section. In each simulation 100 data sets are sampled from the same model. The simulated series are completely comparable since the same error terms are used when generating the data.

Figure 8 presents a graphical summary of the results. We observe that for the linear interpolation predictions the RASE moves further away from the error standard deviation when the error variance increases. Moreover, the increase in spread of RASE values is quite substantial. The penalized spline smoothing predictions result in more stable RASE values across and within simulation settings. For the smallest error variance (0.2^2) the performances are comparable, but overall the penalized spline criterion outperforms linear interpolation in all considered simulation settings.

Also in Figure 8, the stability of the method with respect to the number of selected components is explored. It should be remarked that we do not necessarily have a target number of two components, since the sequentially built warping function might need more components than the actual true warping function. For the penalized spline criterion the selected number of components remains rather constant throughout the simulation (except for $\sigma = 0.2$), while for linear interpolation the results are more variable. For the smallest error variance again the smallest difference between the methods is observed and both are less stable concerning the number of components they select. However, this is because less of the phase variation can be accounted for by the error term and the true underlying warping function is better approximated.

We conclude that for small error variances the linear interpolation fitting criterion performs almost equally well as the penalized spline one, while the penalized spline criterion comes at higher computational cost. This makes us favor linear interpolation in case of little variable error terms. For larger error variances the penalized spline approach is preferred since it offers a more stable, in general better and often faster alignment.

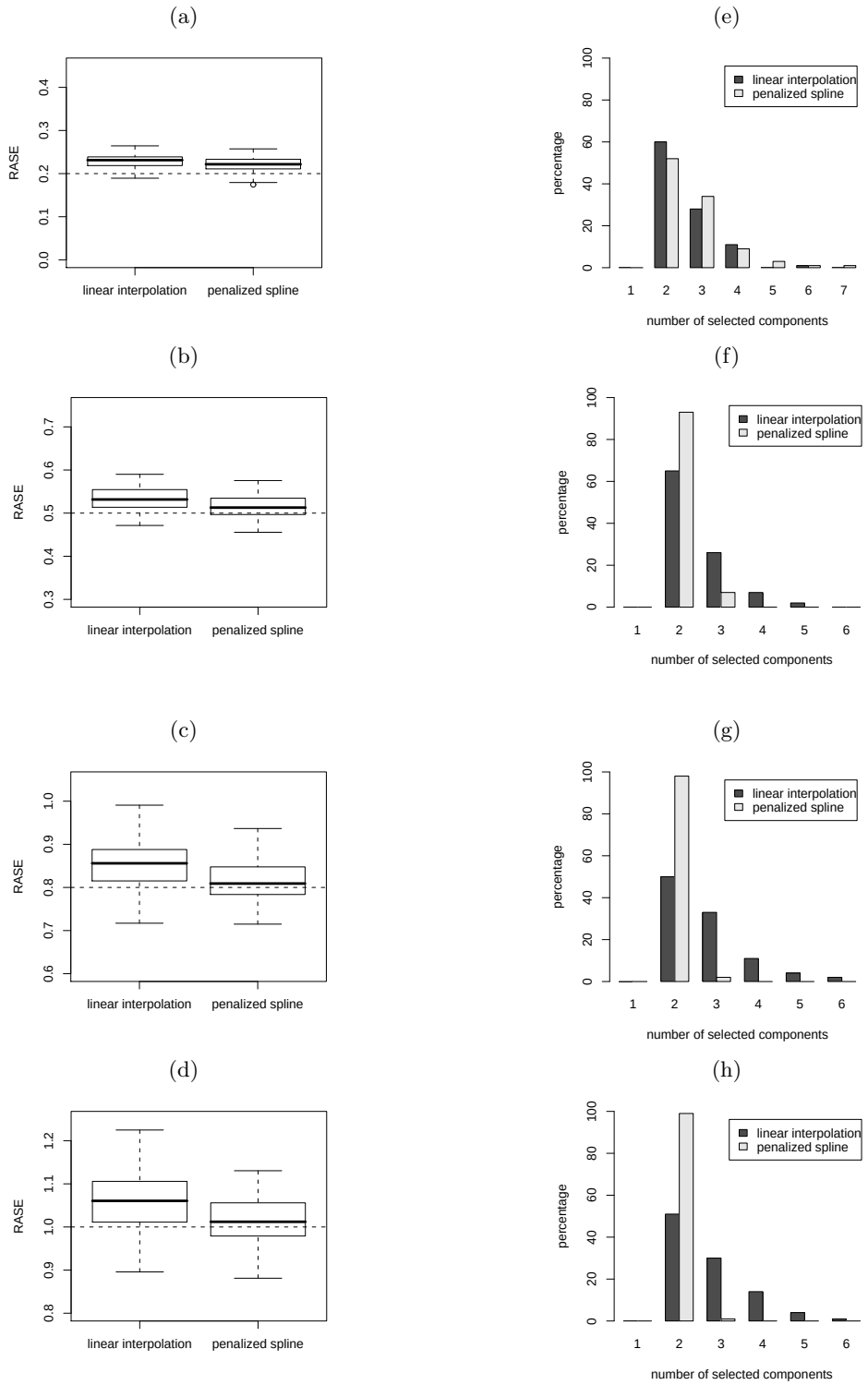


Figure 8: (a)-(d): boxplots of the RASE for an error variance of respectively 0.2^2, 0.5^2, 0.8^2 and 1 (top to bottom). (e)-(h): barplots of the number of selected components.

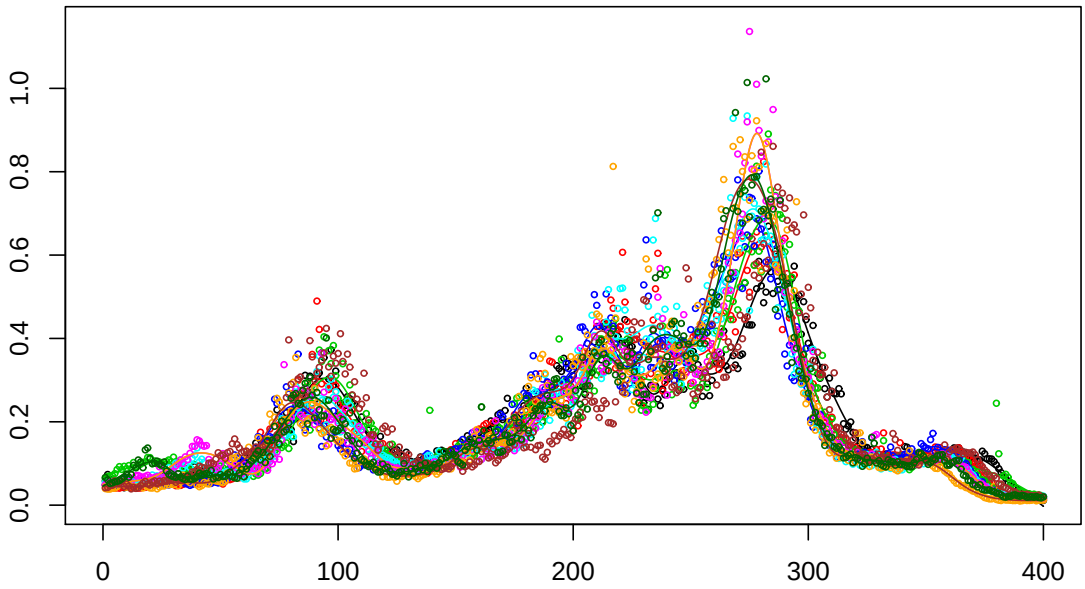
5 Application to the Liquid Chromatography - Mass Spectrometry data set

Proteins play a crucial role in the biological processes guided by a person's DNA towards actual activities in the body. The presence of certain protein patterns could indicate particular normal events, but also diseases or their development, and is therefore of high interest in the field of medical diagnosis and treatment. In this section we will provide an application to a simplified Liquid Chromatography - Mass Spectrometry (LC-MS) data set, previously analyzed by Listgarten et al. (2005), which potentially contains useful information on the protein content of some mixture. We refer to that paper for more information on LC-MS, while we just provide a brief summary. A mass spectrometer is a technological tool that can generate a spectrum based on a protein mixture. In this procedure the sample is split into parts, based on a certain property of the molecules, after which these reduced samples are one by one supplied to the mass spectrometer. Multiple spectra arise, creating a two-dimensional time series spectrum. The latter can be collapsed into a one-dimensional time series, by summing over all the values in the spectrum at each time point, yielding the so-called total ion count (TIC) time series.

The original data set contains 11 such time series, from which 2 were removed for causing too much amplitude variability (Figure 9 (a)). Warping of the time-axis is required to correct for timing differences in the experiment across samples. Our registration approach is applied with the quartic kernel and linear interpolation predictions for a final chain length of 2500 with thinning by 50. As already annotated, it might be of interest to warp all curves to some sort of 'horizontal average'. For two curves one can take two warping functions where one is the inverse of the other, but for multiple curves this is no longer unambiguously generalizable. A possible solution is to put restrictions on the warping parameters. In particular we consider fixed warping bounds ($w_{l,s}$ and $w_{u,s}$) and locations (a_s) for each component, where the only remaining curve specific parameters are the relative intensities $\lambda_{i,s}$, which sum to zero as a 'horizontal averaging' constraint. The sequential Bayesian strategy elapses as described in section 3.3, where the addition of components is stopped when the last component is sparse for all curves (due to the common warping domain or curve-specific $\lambda_{i,s}$).

The estimated warping functions contains only two components and the warped curves are shown in Figure 9 (b). The first component focusses on the area in between time points 180 and 350. It aligns the peaks around 280 to the 'average' peak location, while the second component warps the peak around time point 95. This entire warping stage is summarized by only 22 parameters ($\{\lambda_{i,s}, a_s, w_{l,s}, w_{u,s}\}_{i=1,\ldots,8}^{s=1,2}$). The $\lambda_{i,s}$ provide interesting information concerning the alignment process. For instance, if $\lambda_{i,2} > 0$, the first peak for curve i occurs sooner than the 'average' timing.

(a)



(b)

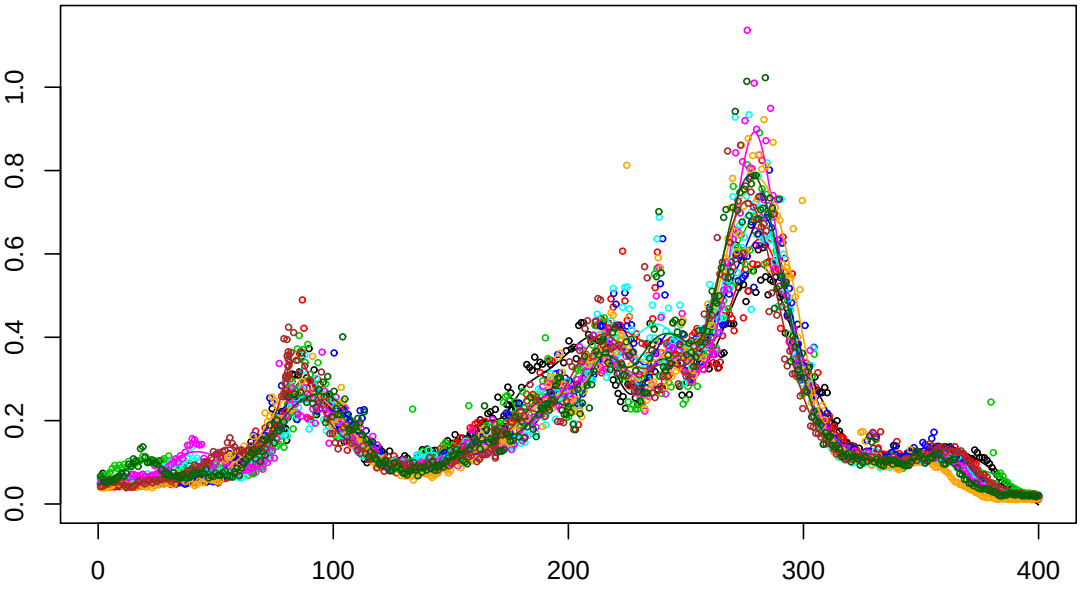


Figure 9: Original curve observations (a) and warped curves (b), including smoothed curves.

(a) (b) (c)

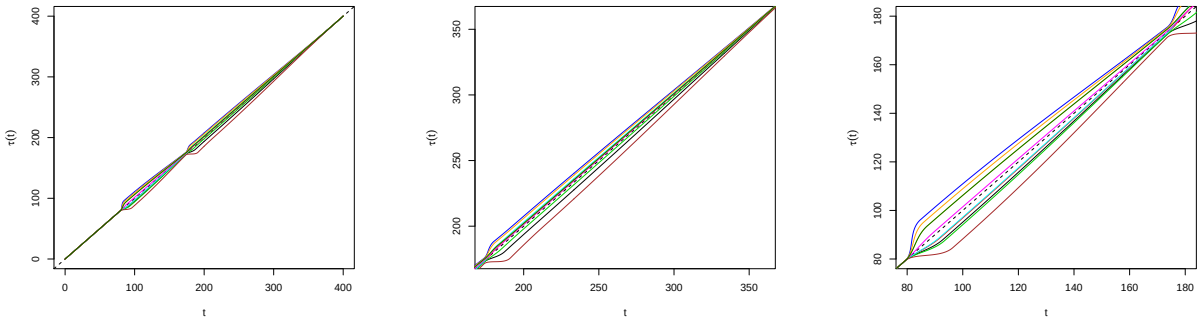


Figure 10: Estimated warping functions (a), zoomed in on the first warping component (b) and on the second warping component (c).

6 Discussion

We introduced the warping component functions as a special type of elementary functions and have explored their properties. Subsequently the need arose for a special fitting strategy in order to estimate such a decomposition structure. This lead to the Bayesian prior-posterior transfer fitting strategy, which incorporates a computationally efficient model selection technique concerning the number of warplets. The obtained decomposition structure summarizes the warping functions by a compact number of interpretable, meaningful components localized in position, scale and intensity, which offer valuable information regarding the curve alignment phase in the analysis. Additionally, the Bayesian framework allows one to conduct exact inferences. Besides the usual MCMC specifications, the method can be considered completely automatic and nonparametric with respect to the warping function.

For more general applications, model (4) needs to be extended. The incorporation of amplitude variability is one of the most important considerations. The current method might still perform well in models with limited amplitude variability (as shown in section 5), however could fail in the presence of stronger amplitude variability.

An interesting extension of the method is to consider images instead of curves. Two-dimensional warplets can be used, for example, for the analysis of grey-scale images. First define the rotation

$$R_{(a,b),\theta}(x,y) = \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x-a \\ y-b \end{pmatrix}.$$

Next we define a function τ_i^x that deforms the surface parallel to the horizontal axis,

$$\tau_i^x\left((a,b,r,\lambda);(x,y)\right) = \begin{pmatrix} \tau_{i,x}^x\left((a,b,r_x,r_y,\lambda);(x,y)\right) \\ y \end{pmatrix},$$

$$\tau_{i,x}^x\left((a,b,r_x,r_y,\lambda);(x,y)\right) = \begin{cases} a + r_x^y \cdot g\left[cK(y)r_x/r_x^y\lambda;(x-a)/r_x^y\right], & x \in [a-r_x^y, a+r_x^y] \\ x, & \textit{otherwise}, \end{cases}$$

with $r_x^y = \sqrt{r_x^2 - (y-b)^2 r_x^2/r_y^2}$ and g, K and λ as in Definition 2.1 and c as in Table 1. We now use the rotation function R to rotate the warping function τ_i^x to make it operate in an arbitrary direction. This defines the two-dimensional warplet as

$$\tau_i\left((a,b,r,\lambda,\theta);(x,y)\right) = \begin{pmatrix} \tau_{i,x}\left((a,b,r_x,r_y,\lambda,\theta);(x,y)\right) \\ \tau_{i,y}\left((a,b,r_x,r_y,\lambda,\theta);(x,y)\right) \end{pmatrix} = R_{(a,b),\theta}\left[\tau_i^x\left((a,b,r_x,r_y,\lambda);R_{(a,b),\theta}^{-1}(x,y)\right)\right].$$

For τ_i^x, the warping interval $[a-r, a+r]$ is extended in a natural way towards an elliptical disk with center (a, b) and half-axes r_x and r_y. One can imagine this component as consisting of one-dimensional warplets parallel to the X-axis with a decreasing warping domain when the y-coordinate moves away from b. The intensities of these warplets are tapered by means of the warplet kernel K, to achieve the same warping behaviour in both the X and Y direction. Apart from having more parameters for each component, we do not expect any fundamental difficulties for the generalization of the fitting procedure to the two-dimensional case.

The decomposition structure obtained from each curve contains important information which can be used for clustering or prediction at a later stage of the analysis. James (2007) applies principal components on the registered curves as well as the warping functions after putting both on a comparable scale. The direct comparability of the λ in a multiple curves setting, as annotated above, offers a meaningful dimension reduction, and makes the warping functions easy to incorporate in existing methods. For example, the principal component analysis for mixed data by Ramsay and Silverman (2006) applied to warping by time shifts could be extended towards our more flexible warping functions.

A Appendix

A.1 Theoretical properties and proofs of section 2.2

Proof of Theorem 2.1

The associativity of the composition in $\mathcal{W}_K$ is obvious since it holds for arbitrary functions.
The neutral element of $\mathcal{W}_K$ is $\epsilon \equiv I$, the identity function on $[l, u]$. It is an element of $\mathcal{W}_K$ since $I \equiv \tau_i(a, 0, r)$ and $\tau \circ I = \tau = I \circ \tau,\ \forall \tau \in \mathcal{W}_K$.
Finally, every element τ in $\mathcal{W}_K$ should have an inverse τ^{-1}. We demonstrate this for a warping component from which it can easily be generalized towards any finite composition. Consider $\tau_i(a, \lambda, r)$ and take an arbitrary $t \in [a-r, a+r]$. We have

$$\tau_i(a,\lambda,r) \circ \tau_i(a,-\lambda,r)(t) \quad = \quad a + rg\left(\lambda, \frac{\tau_i\left((a,-\lambda,r);t\right) - a}{r}\right) = a + r\left(z + \lambda K(z)\right)$$

with $\left(\tau_i\left((a,-\lambda,r);t\right) - a\right)/r = z - \lambda K(z)$. Since

$$\frac{\tau_i\left((a,-\lambda,r);t\right) - a}{r} = \frac{\left[a + r\left(z^* - \lambda K(z^*)\right)\right] - a}{r}, \text{ with } \frac{t-a}{r} = z^* + \lambda K(z^*)$$

it follows that $z = z^*$ and hence $[\tau_i(a, \lambda, r) \circ \tau_i(a, -\lambda, r)(t) = a + r\left(z^* + \lambda K(z^*)\right) = t$. ■

The group of warping functions models a strictly monotone transformation of the dependent variable t of some function F_i displaying phase variability. We will rule out shifts or other transformations which alter the domain of the curves. The latter can always be applied in a pre-processing step. Thus without loss of generality, we make the assumption that there exist real unobserved strictly monotone increasing (s.m.i.), surjective, continuous transformations

$$m_i \in \mathcal{M} = \{\textit{strictly monotone increasing, surjective, continuous functions } m : [l,u] \to [l,u]\},$$

for which $(m_i(t), F_i(t))$ are the registered curves and $(m_i(t_{ij}), f_{ij})$ the registered curve observations. When modelling the unknown m_i by candidates in $\mathcal{W}_{K,[l,u]}$, it is desirable that the warping functions can approximate the more general functions m_i close enough, using distance function d:

$$d : \mathcal{M} \to \mathcal{M}, \textit{ with } d(m_1, m_2) = \sup_{x \in [l,u]} |m_1(x) - m_2(x)| .$$

Theorem A.1 *Take an arbitrary x and y in (l,u). There exists a warping function τ in $\mathcal{W}_K$ that can warp x to y ($\tau(x) = y$). Moreover one can choose τ such that $\tau(t) = t$ for $t \leq \min(x-\varepsilon, y-\varepsilon)$ and $t \geq \max(x+\varepsilon, y+\varepsilon)$, with $\varepsilon > 0$, arbitrarily small.*

Proof of Theorem A.1

In case $x = y$, simply take a warping component with $\lambda = 0$. In what follows we assume $x \neq y$.

For $\mathcal{W}_t$ $(K = K^t, c = 1)$ we can take a single warping component function $\tau = \tau_1$ with warping parameters (a_1, λ_1, r_1) as described in scheme 1 below for $i = 1$, $y_1 = y$ and $x_1 = x$. The idea is to choose the warping parameters such that $|x - \tau_1(x)| > |t - \tau_1(t)|$, $\forall t \in [l,u]$. This is achieved by taking $x = a_1 - c\lambda_1 r_1$ and $y = \tau_1(x) = a_1 + c\lambda_1 r_1$ (which makes (x,y) the top of the rescaled, rotated warplet kernel in τ_1).

Scheme 1 ($l < x \leq x_i < y_i \leq y < u$ or $l < y \leq y_i < x_i \leq x < u$)

$$\begin{aligned} a_i &= (y_i + x_i)/2, \quad \lambda_i = (a_i - x_i)/(cr_i) \\ r_i &= |a_i - x_i| + \varepsilon_i = |a_i - y_i| + \varepsilon_i, \;\; 0 < \varepsilon < min(x-l, y-l, u-x, u-y). \end{aligned}$$

We justify this scheme for the component parameters. Clearly $r_i > 0$. Next, suppose $\min(x-l, y-l, u-x, u-y) = x - l$, this means that $a_i > x$, hence $r_i < a_i - x_i + x - l \leq a_i - l$, thus $a_i - r_i > l$. Also $r_i < y_i - a_i + u - y \leq u - a_i$ implying $a_i + r_i < u$. For the cases in which $\min(x-l, y-l, u-x, u-y)$ equals $y-l, u-x$ or $u-y$, the same can be easily obtained. We still need to check $|\lambda_i| < 1$. Suppose

$x_i < a_i$ (and thus $y_i > a_i$) it then holds that $\frac{(x_i-a_i)}{c|a_i-x_i|} < \lambda_i < \frac{(a_i-x_i)}{c|a_i-x_i|}$ which implies $-1 < \lambda_i < 1$ in $\mathcal{W}_t$. The same can be obtained for $x_i > a_i$ and $y_i < a_i$.

We conclude that τ_1 is indeed a warping component and hence τ a warping function of $\mathcal{W}_t$. Choosing $\varepsilon_1 = \varepsilon$ proves the statement in the lemma.

Finally we have that since $\lambda_1 = (a_1 - x)/(cr_1)$

$$x = a_1 - c\lambda_1 r_1 \text{ and } y = 2a_1 - x = 2a_1 - a_1 + c\lambda_1 r_1 = a_1 + c\lambda_1 r_1 = \tau_1(a_1 - c\lambda_1 r_1) = \tau_1(x).$$

For the other two kernels, a single component will not always suffice. It follows from scheme 1 that since $x_i = a_i - c\lambda_i r_i$ and $(a_i - r_i) > l$ is required, one needs to have

$$x_i = a_i - c\lambda_i r_i > (a_i - r_i)c + a_i(1-c) > lc + a_i(1-c) = a_i - (a_i - l)c = ll + (a_i - l)(1-c).$$

For y_i we obtain $y_i < u-(u-a_i)(1-c)$ and together this is $l+(a_i-l)(1-c) < x_i < y_i < u-(u-a_i)(1-c)$ for $x_i < y_i$. When $x_i > y_i$ these conditions alter towards $l+(a_i-l)(1-c) < y_i < x_i < u-(u-a_i)(1-c)$. This implies that $|y_i - x_i| < c(u-l)$, which also follows from

$$|y_i - x_i| = |a_i - c\lambda_i r_i - (a_i + c\lambda_i r_i)| = |2c\lambda_i r_i| < |2cr_i| < |2c(u-l)/2| < c(u-l).$$

However these conditions are not equivalent.

The construction of the desired warping function is based on scheme 1 for a series of components τ_i with different x_i and y_i values. We consider the situation $x \le l + ((x+y)/2 - l)(1-c) < y$. Note that we can never have $x, y \le l + ((x+y)/2 - l)(1-c)$ or $x, y \ge u - (u - (x+y)/2)(1-c)$. Since we can not apply scheme 1 directly for $x_1 = x$ and $y_1 = y$, we aim at warping x first by a component τ_2 towards a certain value $\tau_2(x)$ as such that $x_1 = \tau_2(x)$ and $y_1 = y$ satisfy the conditions of scheme 1. To this purpose consider the scheme for τ_2 with $y_2 = (x + ny)/(n+1)$ (n a number) and $x_2 = x$, thus $a_2 = ((n+2)x + ny)/2(n+1)$. In what follows we always take $\varepsilon_i = \varepsilon$. We would then have in $\mathcal{W}_e$ and $\mathcal{W}_q$ ($K = K^e, c = 1/2$ and $K = K^q, c = (3\sqrt{3})/8$ resp.) that

$$\frac{2n+3}{2n+2}y < u + \frac{x}{2n+2} \Rightarrow u - (u - a_1)(1-c) > u - \left(u - \frac{x + (2n+1)y}{2(n+1)}\right)\frac{1}{2} > y. \tag{11}$$

In case $y < u - (u - (x+y)/2)(1-c)$ we can take $n = 0$, otherwise there exists a certain n_1 as such that (11) holds for $n = n_1$. Indeed $\lim_{n\to\infty} \frac{2n+3}{2n+2}y = y < u = \lim_{n\to\infty} u + \frac{x}{2n+2}$ and from now on we

consider n to be equal to n_1. For the other inequality it is clear that

$$\begin{aligned} l + (a_1 - l)(1 - c) &\leq \frac{1}{2}l + \left(\frac{x + (2n+1)y}{2(n+1)}\right)\frac{1}{2} \\ &< \frac{1}{2}\left(\frac{3x}{2(n+1)} + \frac{(2n-1)y}{2(n+1)} + \frac{x}{2(n+1)} + \frac{(2n+1)y}{2(n+1)}\right) \\ &= \frac{1}{2}\left(\frac{2x}{(n+1)} + \frac{2ny}{(n+1)}\right) = \frac{x+ny}{(n+1)} = y_2 = \tau(x_2) = x_1. \end{aligned} \tag{12}$$

Thus applying τ_2 first would solve the problem for τ_1. Next we verify the conditions of scheme 1 for τ_2. Since $y_2 < y_1 = y$ we have lowered the barrier for $x_2 = x$ of scheme 1 towards $l + (a_2 - l)(1 - c)$. However, it might still be possible that $x > l + (a_2 - l)(1 - c)$. This would require another component τ_3 with $y_3 = (x + ny_2)/(n + 1)$ and $x_3 = x$, and for τ_2 and τ_1 this means $x_2 = \tau_3(x)$, y_2 as before and $x_1 = (x + ny)/(n + 1) = \tau_2 \circ \tau_3(x)$, $y_1 = y$. If we continue on this way, we create a sequence of components τ_i following scheme 1 for

$$x_i = ax\sum_{j=0}^{i-1} b^j + b^i y, \quad i > 2, \quad \text{with } a = \frac{1}{n+1} \text{ and } b = \frac{n}{n+1}\ , \tag{13}$$

$$y_i = ax\sum_{j=0}^{i-2} b^j + b^{i-1} y, \quad i > 2, \tag{14}$$

yielding $a_i = ax\sum_{j=0}^{i-2} b^j + 0.5axb^{(i-1)} + 0.5b^{(i-1)}y(1 + b), i > 2$. We now have

$$\lim_{i\to\infty}\frac{x + y_i}{2} = \frac{x}{2} + \frac{1}{2}\frac{x}{(n+1)}\frac{1}{\left(1 - \frac{n}{n+1}\right)} + \lim_{i\to\infty}\left(\frac{n}{n+1}\right)^{(i-2)} y = x.$$

The fact that $x > l + (x - l)(1 - c)$ guarantees the existence of a value k_1 such that $x > l + ((x + y_i)/2 - l)(1 - c)\ \forall\ i \geq k_1$. Thus we can apply the scheme for $x_i = x$ and y_i as in (14) $\forall\ i \geq k_1$. It is easy to see that $\lim_{i\to\infty} y_i = x < 1/2u + x \leq \lim_{i\to\infty} u - (u - (x + y_i)/2)(1 - c)$. This proves that there is a k_2 such that the second condition of scheme 1 holds for $y_i\ \forall\ i \geq k_2$ We can thus take the warping function $\tau = \tau_1 \circ \ldots \circ \tau_{k-1} \circ \tau_k$ for $k = \max(k_1, k_2)$, with component parameters (a_1, λ_1, r_1) as set out by scheme 1 with x_i and y_i as follows:

$$\begin{aligned} &x_1 = (x + ny)(n + 1) \text{ and } y_1 = y, \quad x_k = x \text{ and } y_k \text{ as in equation (14)}, \\ &x_i \text{ and } y_i \text{ as in equations (13) and (14) resp. for } i = 2, \ldots, k - 1. \end{aligned}$$

Finally we prove that the conditions of scheme 1 are satisfied for the components $\tau_2, \ldots, \tau_{k-1}$. We start with $y_i < u - (u - a_i)(1 - c)$ by means of induction.

- *i=2.* We know that for y_i and $x_i = (x + ny_i)/(n+1)$ the condition can be rewritten as in (11). Thus for $y_2 = (x + ny)/(n+1)$ and $x_2 = (x + ny_2)/(n+1)$ we need to show that

$$\begin{aligned}
\frac{2n+3}{2n+2}y_2 < u + \frac{x}{2n+2} &\Leftrightarrow \frac{2n+3}{2n+2}\left(\frac{1}{n+1}x + \frac{n}{n+1}y\right) < u + \frac{x}{2n+2} \\
&\Leftrightarrow \frac{2n+3}{2n+2}y + \frac{n+1}{n(n+1)}x < \frac{n+1}{n}u + \frac{n+1}{n(2n+2)}x \\
&\Leftrightarrow \frac{2n+3}{2n+2}y < u + \frac{1}{(2n+2)}x + \frac{1}{n}u - \frac{2n+1}{n(2n+2)}x.
\end{aligned}$$

  Because of (11), the above holds if $\frac{1}{n}u - \frac{2n+1}{n(2n+2)}x \geq 0$. This would be true when $\frac{1}{n} \geq \frac{2n+1}{n(2n+2)}$, since $u > x$. The latter follows from $n \geq 0$.

- *Induction step.* Assume that the condition holds for $i \geq 2$ and we will show that it also holds for $i+1$. Take $y_{i+1} = (x + ny_i)/(n+1)$, $x_{i+1} = (x + ny_{i+1})/(n+1)$ and repeat the previous reasoning to obtain:

$$\begin{aligned}
\frac{2n+3}{2n+2}y_{i+1} < u + \frac{x}{2n+2} &\Leftrightarrow \frac{2n+3}{2n+2}y_i < u + \frac{1}{(2n+2)}x + \frac{1}{n}u - \frac{2n+1}{n(2n+2)}x \\
&\Leftarrow \frac{1}{n}u - \frac{2n+1}{n(2n+2)}x > 0 \\
&\Leftarrow \frac{1}{n} > \frac{2n+1}{n(2n+2)} \Leftrightarrow n > 0.
\end{aligned}$$

For the second condition $x_i > l - (l - a_i)(1 - c)$ we can proceed as in (12),

$$l + (a_i - l)(1-c) \leq \frac{1}{2}l + \left(\frac{x_i + (2n+1)y_i}{2(n+1)}\right)\frac{1}{2} < \frac{x + ny_i}{(n+1)} = x_i. \qquad \blacksquare$$

Theorem A.2

1. *The set $\mathcal{W}_K$ is a subset of $\mathcal{M}$ and the pair $(\mathcal{M}, d)$ is a metric space.*

2. *$\mathcal{W}_K$ is bounded and its diameter $\delta(\mathcal{W}_K) = \sup_{\tau_1,\tau_2 \in \mathcal{W}_K} d(\tau_1, \tau_2) = (u - l)$.*

In the framework of the metric space $(\mathcal{M}, d)$:

3. *$\mathcal{W}_K$ has no internal points: $\mathring{\mathcal{W}}_K = \emptyset$.*

4. *$\mathcal{W}_K$ has no external points.*

5. *The edge of $\mathcal{W}_K$ is equal to the entire space $\mathcal{M}$: $\partial\mathcal{W}_K = \mathcal{M}$.*

6. *$\mathcal{W}_K$ is dense in $\mathcal{M}$, that is $\overline{\mathcal{W}}_K = \mathcal{M}$.*

Proof of Theorem A.2

1. Since every warping component τ_i is continuous and s.m.i., so is the composition τ.

We know that $\tau(t) = t$ for every $t \in [l, u] \setminus \{\cup_{i=1,\ldots,n}[a_i - r_i, a_i + r_i]\}$. Thus if $a_i - r_i > l$ and $a_i + r_i < ul$, $\forall i$, then $\tau(l) = l$ and $\tau(u) = u$, In case some of the $a_i - r_i = l$ and/or some of the $a_i + r_i = u$, then for those i: $\tau_i(l) = l$ and/or $\tau_i(u) = u$ and thus also $\tau(l) = l$ and $\tau(u) = u$. Since τ is a continuous function, all function values between l and u must be attained on $[l, u]$ and hence τ is surjective.

Also, we easily obtain that $\forall h_1, h_2, h_3 \in \mathcal{H}$:

$$\begin{aligned}
d(h_1, h_2) &= \max_{t\in[l,u]} |h_1(t) - h_2(t)| \geq 0 \\
d(h_1, h_2) &= \max_{t\in[l,u]} |h_1(t) - h_2(t)| = \max_{t\in[l,u]} |h_2(t) - h_1(t)| = d(h_2, h_1) \\
d(h_1, h_2) &= \max_{t\in[l,u]} |h_1(t) - h_2(t)| = \max_{t\in[l,u]} |h_1(t) - h_3(t) + h_3(t) - h_2(t)| \\
&\leq \max_{t\in[l,u]} (|h_1(t) - h_3(t)| + |h_3(t) - h_2(t)|) = d(h_1, h_3) + d(h_3, h_2). \\
d(h_1, h_2) = 0 &\Leftrightarrow \forall t \in [l, u] : |h_1(t) - h_2(t)| = 0 \Rightarrow h_1 = h_2.
\end{aligned}$$

2. We have that $\forall \tau \in \mathcal{W}_K$ and $\forall t \in [l, u] : l \leq \tau(t) \leq u$. Thus

$$\delta(\mathcal{W}_K) = \sup_{\tau_1,\tau_2\in\mathcal{W}_K} (d(\tau_1, \tau_2)) = \sup_{\tau_1,\tau_2\in\mathcal{W}_K} \left(\max_{t\in[l,u]} |\tau_1(t) - \tau_2(t)| \right) \leq u - l.$$

For the other inequality we need to show that:

$$\forall \varepsilon > 0 : \ \exists \tau_1, \tau_2 \in \mathcal{W}_K \text{ with } \max_{t\in[l,u]} |\tau_1(t) - \tau_2(t)| > l - u - \varepsilon.$$

Thus take an arbitrary, fixed $0 < \varepsilon < (u - l)/3$. We then select the following two warping functions in $\mathcal{W}_K$:

- τ_1 a warping function that warps $x = (l + u)/2$ towards $y = u - \varepsilon/3$.
- τ_2 a warping function that warps $x = (l + u)/2$ towards $y = l + \varepsilon/3$.

Lemma A.1 guarantees the existence of these warping functions. For these choices it holds that

$$|\tau_1(x) - \tau_2(x)| = \left| (u - \frac{\varepsilon}{3}) - (l + \frac{\varepsilon}{3}) \right| = u - l - \frac{2}{3}\varepsilon > u - l - \varepsilon.$$

3. We need to prove that for every $\tau \in \mathcal{W}_K$, and for every $r > 0$, there exists a function m such that $m \in B_r(\tau) \cap (\mathcal{M} \setminus \mathcal{W}_K)$. Take an arbitrary $\tau \in \mathcal{W}_K$, and denote the number of components by K. Since τ is continuous, there exists a point $c_r \in [l, u]$, with $\tau(c_r) = u - r$. This c_r is smaller than u, since τ is s.m.i. Now consider the function m defined by $m(t) = \tau_{K+1} \circ \tau(t), \ \ \forall t \in [l, u]$, where the

warplet τ_{K+1} has warping parameters $(a_{K+1}, \lambda_{K+1}, r_{K+1}) = (\frac{u+c_r}{2}, 0.5, \frac{u-c_r}{3})$ and, importantly, has a different kernel K than the one of the group $\mathcal{W}_K$. In particular for $\mathcal{W}_q$ take $K = K^t$ and for $\mathcal{W}_e$ and $\mathcal{W}_t$ take $K = K^q$. Clearly this implies that m no longer belongs to $\mathcal{W}_K$. Indeed, for $\mathcal{W}_t$ the warping functions are all piecewise linear while m is not. In $\mathcal{W}_q$ all τ have a first derivative that exist on the entire $[l, u]$ domain, which is not true for m. Finally for $\mathcal{W}_e$ an arbitrary $\tau \neq I$ can never approach ul in the way m does since the largest $a_i + r_i$ of the components is always a location where the first derivative is not defined, while this is not the case for m. On the other hand it is clear that m is a s.m.i., continuous, surjective function on $[l, u]$ and consequently $m \in \mathcal{M} \setminus \mathcal{W}_K$.

Since $d(\tau, m) = 0$, $\forall t \in [l, u] \setminus [c_r, u]$ we have that

$$d(\tau, m) = max_{t\in(c_r,u)} |\tau(t) - m(t)| < ul - \tau(c_r) \leq u - (u - r) = r,$$

since $\tau(c_r) < \tau(t) < u$ and $\tau(c_r) < m(t) < u, \quad \forall t \in (c_r, u)$. Thus m is contained in $B_r(\tau)$ and m is an element of $\mathcal{M} \setminus \mathcal{W}_K$. We conclude that $\mathcal{W}_K$ is open in $\mathcal{M}$.

4. We need to prove that for every $m \in \mathcal{M}$, and for every $r > 0$ there exist a function $\tau \in \mathcal{W}_K$ such that $\tau \in B_r(m)$. Take an arbitrary $m \in \mathcal{M}$ and $0 < r < u - l$. Since $m \in \mathcal{M}$, there exist points k_i $(i = 1, \ldots, n)$ in $[l, u]$ which are described by

$$p = \text{ceil}\,[(u - l)/(r/2)] \quad \text{and} \quad m(k_i) = l + (i - 1)\frac{u - l}{p}, \quad i = 1, \ldots, p + 1.$$

Because of the surjectivity of m these k_i exist and they are distinct because m is s.m.i. In between the points k_i and k_{i+1}, the function m increases by

$$\begin{aligned} m(k_{i+1}) - m(k_i) &= l + (i)\frac{u - l}{p} - \left[l + (i - 1)\frac{u - l}{p}\right] = \frac{u - l}{p} \\ &\leq \frac{u - l}{(u - l)/(r/2)} = r/2 < r. \end{aligned} \tag{15}$$

Now consider the warping function $\tau = \tau_{p-1} \circ \ldots \circ \tau_2$, in which $\tau_2, \ldots, \tau_{p-1}$ are also warping functions determined by means of Lemma A.1 such that for τ_i

$$x = \tau_{i-1} \circ \ldots \circ \tau_2(k_i),\ y = m(k_i),\ \varepsilon < \min\left(x - m(k_{i-1}),\ u - x\right), \quad i = 2, \ldots, p - 1.$$

The warping function τ_i is responsible for warping k_i to $m(k_i)$, while not re-warping the previously warped $k_2, \ldots, k_{i-1}$ because of the choice of ε. The latter is possible since $k_{i-1} < k_i$ and thus $m(k_{i-1}) = \tau_{i-1} \circ \ldots \circ \tau_2(k_{i-1}) < \tau_{i-1} \circ \ldots \circ \tau_1(k_i) = x$. The composition of these warping functions τ has the following properties:

$$\tau(k_i) = m(k_i) = l + (i - 1)\frac{u - l}{r - 1}, \quad i = 1, \ldots, p + 1, \tag{16}$$

$$\tau(k_{i+1}) - \tau_r(k_i) < r, \quad i = 1, \ldots, p. \tag{17}$$

We now have that $d(m,\tau) = max_{t\in[l,u]}\,|m(t)-\tau(t)| = max_{t\in\cup_{i=1,\ldots,p}(k_i,k_{i+1})}\,|m(t)-\tau(t)| < r$, because of (15), (17) and (16) and since τ and m are s.m.i.

5. We have shown that every $m \in \mathcal{M}$ is not an external point of $\mathcal{W}_K$. Also every $\tau \in \mathcal{W}_K$ is not an internal point of $\mathcal{W}_K$. Thus all $m \in \mathcal{M}$ are neither internal (elements $m \in \mathcal{M} \setminus \mathcal{W}_K$ are never internal points) nor external points of $\mathcal{W}_K$ and are consequently all edge-points.

6. We easily obtain from the previous that $\overline{\mathcal{W}_K} = \mathring{\mathcal{W}}_K \cup \partial\mathcal{W}_K = \emptyset \cup \mathcal{M} = \mathcal{M}$. ■

References

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*(4), 327–335.

Gervini, D. and T. Gasser (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika 92*(4), 801–820.

James, G. (2007). Curve alignment by moments. *The Annals of Applied Statistics 1*(2), 480–501.

Kneip, A. and T. Gasser (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics 20*(3), 1266–1305.

Listgarten, J., R. M. Neal, S. T. Roweis, and A. Emili (2005). Multiple alignment of continuous time series. In L. K. Saul, Y. Weiss, and L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17*, pp. 817–824. Cambridge, MA: MIT Press.

Park, J. (2008). On the effect of curve alignment and functional PCA. In S. Dabo-Niang and F. Ferraty (Eds.), *Functional and Operatorial Statistics*, pp. 243–249. Physica-Verlag, Heidelberg.

Ramsay, J. and X. Li (1996). Curve registration. *Journal of the Royal Statistical Society, Series B 60*(2), 351–363.

Ramsay, J. and B. Silverman (2006). *Functional Data Analysis* (2 ed.). Spinger, New York.

Silverman, B. (1995). Incorporating parametric effects into functional principal components analysis. *Journal of the Royal Statistical Society, Series B 57*(4), 673–689.

Telesca, D. and L. Inoue (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association 103*(481), 328–339.

Supplemental Material: Topological Concepts

Definition A.1 *The pair (X, d), is called a metric space, if X is a set and d a function $d : X \to X$ which satisfies the following properties, for all $x, y, z \in X$:*

$$d(x, y) \geq 0,\ d(x, y) = d(y, x),\ d(x, y) = d(x, z) + d(z, y),\ d(x, y) = 0 \Rightarrow x = y.$$

Definition A.2 *Consider a metric space (X, d), $A \subset X \neq \emptyset$ and $x \in X$.*

- *the distance between x and A: $d(x, A) = \inf_{a \in A} d(x, a)$*
- *the distance between X and A: $d(A, X) = \inf_{x \in X, a \in A} d(x, a)$*
- *the diameter of A: $\delta(A) = \sup_{a_1, a_2 \in A} d(a_1, a_2)$*
- *A is called to be bounded if: $\delta(A) < \infty$*
- *the open ball with radius r and center a is defined by: $B_r(a) = \{x \in X \mid d(x, a) < r\}$.*

Definition A.3 *Consider a metric space (X, d), $A \subset X \neq \emptyset$ and $x \in X$,$a \in A$.*

- *a is an internal point of A if $(\exists r > 0) : B_r(a) \subset A$,*
- *x is an external point of A if $(\exists r > 0) : B_r(x) \subset (X/A)$,*
- *x is an edge-point of A if x is neither an internal nor external point of A, or if $(\forall r > 0) : B_r(x) \cap A \neq \emptyset$ and $B_r(x) \cap (X \setminus A) \neq \emptyset$.*
- *x is a closure point of A if $(\forall r > 0) : B_r(x) \cap A \neq \emptyset$.*
- *the internal of A, $\mathring{A}$, is the set of all internal points of A*
- *the edge of A, ∂A, is the set of all edge-points of A*
- *the closure of A, $\overline{A}$, is the set of all closure points of A. Or $\overline{A} = \mathring{A} \cup \partial A$.*
- *A is open if every point of a is an internal point, $A = \mathring{A}$, or if:*

$$(\forall a \in A)\ (\exists r > 0) : B_r(a) \subset A,$$

- *A is closed if every point is a closure point, $A = \overline{A}$, or if $(X \setminus A)$ is open:*

$$(\forall x \in (X \setminus A)\ (\exists r > 0) : B_r(x) \subset (X \setminus A),$$

Definition A.4 *A subset A of a metric space (X, d) is called dense in X, if $\overline{A} = X$.*